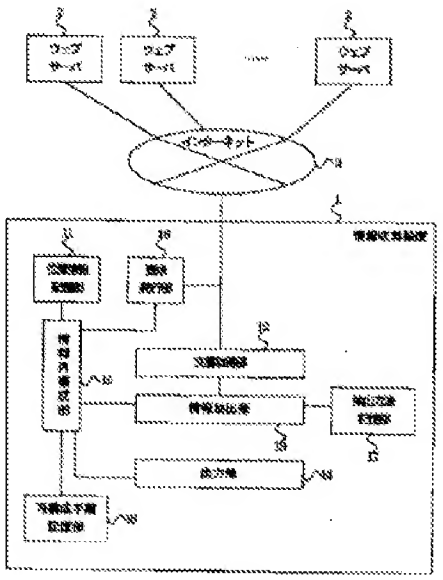
INFORMATION COLLECTION DEVICE AND INFORMATION COLLECTION PROGRAM

Patent number:	JP2003271610
Publication date:	2003-09-26
Inventor:	YAMADA TAKESHI; KASE NAOKI
Applicant:	TOKYO SHIBAURA ELECTRIC CO
Classification:	
- international:	***G06F17/30;*** **G06F17/30;** (IPC1-7): G06F17/30
- european:	
Application number:	JP20020075412 20020319
Priority number(s):	JP20020075412 20020319

Abstract of **JP2003271610**

PROBLEM TO BE SOLVED: To moderate the risk accompanied by information collection with disclosed information on the Internet as an information source and to perform a reliable information collection.
SOLUTION: A plurality of sites for providing intended information is selected from sites disclosed on the Internet. More probable information is selected from respective pieces of information provided by the plurality of sites, reconfigurated and outputted.
COPYRIGHT: (C)2003,JPO



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2003－271610
（P2003－271610A）
(43)公開日　平成15年９月26日(2003.9.26)

(51)Int.Cl.[7]	識別記号	ＦＩ	ﾃｰﾏｺｰﾄﾞ(参考)
Ｇ０６Ｆ　17/30	１８０	Ｇ０６Ｆ　17/30	180Z　5B075
	１１０		110C

審査請求　未請求　請求項の数12　ＯＬ　（全 10 頁）

(21)出願番号　特願2002－75412(P2002－75412)

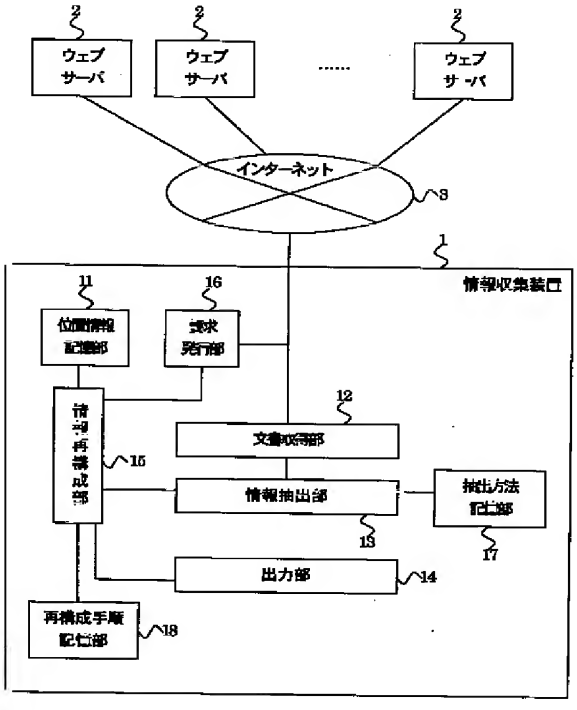
(22)出願日　平成14年３月19日(2002.3.19)

(71)出願人　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(72)発明者　山田　剛
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(72)発明者　加瀬　直樹
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(74)代理人　100083161
弁理士　外川　英明

Ｆターム(参考)　5B075 KK02 PQ36 QS07

(54)【発明の名称】　情報収集装置及び情報収集プログラム

(57)【要約】
【課題】　インターネット上の公開情報を情報源とした情報収集に伴うリスクを緩和し、かつ信頼性の高い情報収集を行なうことを目的とする。
【解決手段】　目的とする情報を提供するサイトを、インターネットに公開されている中から複数選定し、該複数のサイトが提供する情報それぞれから、より確からしい情報を選択し、再構成して出力する。



【特許請求の範囲】
【請求項１】　ネットワークを通じて電子的に公開されている情報をアクセスする情報収集装置であって、
収集対象である情報を含む目的とする情報の所在を示す位置情報について、該収集対象情報について実質的に同一の内容を含む、目的とする情報の所在を示す複数の位置情報を予め記憶しておく位置情報記憶手段と、
一度の取得要求に応じて前記複数の位置情報を用いて実質的に同一の目的とする情報を複数の所在から取得しようとする情報取得手段とを具備したことを特徴とする情報収集装置。
【請求項２】　前記情報取得手段は、前記位置情報記憶手段が記憶したある位置情報から目的とする情報が取得できなかった場合には、前記位置情報記憶手段が記憶する他の位置情報を用いて、該他の位置情報が指し示す別の目的とする情報から収集対象に関する情報を取得しようとすることを特徴とする請求項１に記載の情報収集装置。
【請求項３】　取得した複数の目的とする情報から収集対象情報に関する情報要素を生成する情報生成手段をさらに具備し、前記情報生成手段は、複数の目的とする情報に基づいて得られる情報要素の内から出力として採用するものを、予め与えられた手順に従って選択する選択手段を有するものであることを特徴とする請求項１乃至２に記載の情報収集装置。
【請求項４】　前記情報取得手段は、一度の取得要求に応じて、複数の収集対象情報に関する情報をまとめて取得することを特徴とする、請求項１乃至３に記載の情報収集装置。
【請求項５】　前記情報生成手段によって選択された前記情報要素を、予め指定した出力様式に従って構成し直す情報構成手段をさらに具備することを特徴とする、請求項４に記載の情報収集装置。
【請求項６】　前記選択手段は、複数の前記情報要素の内から出力として採用するものを選択するとき、位置情報によって示されるそれぞれの情報提供サーバの稼動状況と、位置情報によって示されるそれぞれの目的とする情報の、統計的手法に基づく取得対象情報に関する信頼度との少なくともどちらかの、取得状況によって動的に変更される、あるいは予め与えられる情報に基づいて選択が為されることを特徴とする請求項３乃至５に記載の情報収集装置。
【請求項７】　ネットワークを通じて電子的に公開されている情報をアクセスする情報収集プログラムであって、
一度の取得要求に応じて、
収集対象である情報を含む目的とする情報の所在を示す位置情報について、該収集対象情報について実質的に同一の内容を含む、目的とする情報の所在を示す複数の位置情報から、複数の位置情報を選択する第１のステップと、
前記第１のステップで選択した目的とする情報についての複数の位置情報を用いて、収集対象情報について実質的に同一の複数の目的とする情報を取得する第２のステップとを有することを特徴とする情報収集プログラム。
【請求項８】　前記第２のステップで、ある位置情報から目的とする情報が取得できなかった場合には、複数記憶する他の位置情報を用いて、該他の位置情報が指し示す別の目的とする情報から収集対象に関する情報を取得しようとする第３のステップをさらに有することを特徴とする請求項７に記載の情報収集プログラム。
【請求項９】　複数の目的とする情報に基づいて得られる情報要素の内から出力として採用するものを、予め与えられた手順に従って選択し、取得した複数の該目的とする情報から収集対象情報に関する情報要素を生成する第４のステップをさらに有することを特徴とする請求項７乃至８に記載の情報収集プログラム。
【請求項１０】　前記第２または第３のステップは、一度の取得要求に応じて、複数の収集対象情報に関する情報をまとめて取得することを特徴とする、請求項７乃至９に記載の情報収集プログラム。
【請求項１１】　前記第４のステップによって選択された前記情報要素を、予め指定した出力様式に従って構成し直す第５のステップをさらに有することを特徴とする、請求項１０に記載の情報収集プログラム。
【請求項１２】　前記第４のステップは、複数の前記情報要素の内から出力として採用するものを選択するとき、位置情報によって示されるそれぞれの情報提供サーバの稼動状況と、位置情報によって示されるそれぞれの目的とする情報の、統計的手法に基づく取得対象情報に関する信頼度との少なくともどちらかの、取得状況によって動的に変更される、あるいは予め与えられる情報に基づいて選択が為されることを特徴とする請求項９乃至１１に記載の情報収集プログラム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、公開されている情報の収集装置及びプログラムであって、特に複数の情報源から情報を収集し目的とする情報あるいは情報要素の取得確率を高めるとともに、信頼性が高い情報を収集する情報収集装置及びプログラムに関する。
【０００２】
【従来の技術】インターネットの発達によって電子化されたさまざまな情報を入手することが容易になってきた。昨今では、今までは対価を払って入手していた情報も、インターネットから無償で入手できることが多い。あるいは高価な対価を払って契約する既存の情報提供サービスも、インターネットを介して行なわれることが多くなってきた。インターネット上で得られる情報は、一般にウェブページの形で提供されており、インターネッ

トの利用者はウェブブラウザーを使用してこれらの情報にアクセスしている。

【０００３】これらウェブページを元にした情報収集には大きく二つの問題がある。一つめはHTML(HyperText Markup Language)形式で提供されるために、そこに含まれる情報の二次利用が困難であることである。HTML形式は画面上に視覚的な情報を投影するには好都合であるが、反面、データが整然と表記されたものにはならない。この問題をある程度克服するものとして特開２０００－３４８０６１号公報にあるような、文書構造を解析しデータの抽出およびデータの二次利用を行なうための発明がされている。

【０００４】もう一つは、ワールドワイドに点在する、情報源となるウェブサイトの全てが必ずしも厳格な管理の元に運営されているわけではないことである。提供する情報の信頼性を保証する契約の元に運営されるウェブサイトは別として、インターネット上の無償で提供される有益なウェブサイトのほとんどは、提供する情報の信頼性を保証していない。これは当該ウェブサイトの運営が保証されているものではなく、また必ずしも常に正しい情報を提供しなければならない義務を負っているものでもないことを意味している。このような実情から、情報源となるそれぞれのウェブサイトはウェブサイト運営者の事情による不意の停止や提供ウェブページの変更、あるいは提供情報に誤りが有り得ることを十分考慮しなければならない。上記実情を鑑みれば、定常的にサービスを行なうシステムの情報源として既述のウェブサイトを使用することは困難であり情報更新のリアルタイム性が確保できないばかりでなく、場合によっては自らのサービスの停止を余儀なくされることさえ有り得る。さらには誤った情報に基づいて情報をサービスし続けてしまうことなども懸念される。このことは当該システムを提供する企業等にとって信用失墜などの極めて重大な問題を引き起こし、またそれによる機会損失を蒙るなど、その悪影響は計り知れない。

【０００５】

【発明が解決しようとする課題】インターネット上の公開情報を用いて情報収集を行なう際の上記のような問題点を緩和し、インターネット上の公開情報を情報源とした情報の取得率と提供する情報の信頼性の向上を図ることを目的とする。

【０００６】

【課題を解決するための手段】第１の発明によれば、ネットワークを通じて電子的に公開されている情報をアクセスする情報収集装置であって、収集対象である情報を含む目的とする情報の所在を示す位置情報について、該収集対象情報について実質的に同一の内容を含む、目的とする情報の所在を示す複数の位置情報を予め記憶しておく位置情報記憶手段と、一度の取得要求に応じて前記複数の位置情報を用いて実質的に同一の目的とする情報を複数の所在から取得しようとする情報取得手段とを具備したことを特徴とする情報収集装置が提供される。また、第２の発明によれば、ネットワークを通じて電子的に公開されている情報をアクセスする情報収集プログラムであって、一度の取得要求に応じて、収集対象である情報を含む目的とする情報の所在を示す位置情報について、該収集対象情報について実質的に同一の内容を含む、目的とする情報の所在を示す複数の位置情報から、複数の位置情報を選択する第１のステップと、前記第１のステップで選択した目的とする情報についての複数の位置情報を用いて、収集対象情報について実質的に同一の複数の目的とする情報を取得する第２のステップとを有することを特徴とする情報収集プログラムが提供される。

【０００７】これによって、実質的に同一の情報を複数収集し、継続的な目的とする情報の収集と、より信頼のおける情報の取捨選択とを行なわせることが可能となる。

【０００８】

【発明の実施の形態】第１の実施形態における情報取得装置のシステム構成図の一例を図１に示す。情報収集処理装置１と、実質的に相互に同等な情報を内包する半構造化文書、たとえばHTML(HyperText Markup Language)形式のウェブページを電子的手段によって提供するサーバ、たとえばウェブページを提供するウェブサーバ２と、それが複数接続されたインターネット３とが示されている。

【０００９】次に情報収集装置１について説明する。情報再構成部１５は再構成手順記憶部１８に記憶された文書取得手段に従って、少なくとも１つ以上のウェブページを文書取得部１２に指示して取得させる。そして各々のウェブページに内在する情報要素を、抽出方法記憶部１７に記憶された抽出方法に従って情報抽出部１３に抽出させる。このようにして抽出された情報要素は、情報再構成部１５によって情報の再構成、たとえばHTMLへの再編集あるいはリスト形式のデータに加工されるなどして、出力部１４を介して出力される。

【００１０】図２は情報収集装置１の動作のフローの一例を示す図である。情報再構成部１５は位置情報記憶部１１に記憶されたそれぞれのウェブサーバ２の、たとえばURL(Uniform Resource Locator)を取得する（ステップＳ１）。URLはサーバがインターネット３における位置と、ウェブページの格納場所を示す位置情報の組を示す規定の形式である。続いて当該URLが示すウェブページを取得すべく、要求発行部１６からインターネット３を通じてウェブサーバ２に対して要求する（ステップＳ２）。要求を受け取ったウェブサーバ２はこれに呼応してURLにて指定されたウェブページを情報収集装置１に対して送り返す。情報収集装置１は、この送り返されてきたウェブページを文書取得部１２にて受け取る（ステ

ップＳ３）。このとき、ウェブサーバ２やインターネット３等で不具合が生じない限り、URLで表現したウェブページが取得できる。

【００１１】取得に成功したウェブページはある文書構造（たとえばHTML）にしたがって情報要素（たとえばHTML内の株価情報など）を内包している。各々のウェブページはそれぞれが別個の文書構造をしているのが普通であるから、情報抽出部１３はそれぞれの文書構造に応じた予め記憶してあるデータ取得方法を抽出方法記憶部１７から取得（ステップＳ４）し、その取得方法によって出力となる１組の情報要素群を抽出する（ステップＳ５）。

【００１２】情報要素の抽出に際しては、取得したウェブページが抽出方法記憶部１７に記憶された抽出方法を検討したときと同じものであれば何ら問題は生じない。場合によっては、たとえばウェブサーバ２の運営者によって当該ウェブページを変更した場合や、ウェブサーバ２の不具合によって指示したのとは別のウェブページが取得される場合も考えられる。どのようなウェブページが取得されたかにもよるが、一部あるいは全部の情報要素の取得に失敗する、もしくは誤った情報要素を抽出してしまう。このことは情報収集という目的において大きな問題となる。しかしながらこのような事態の発生は予測が不可能である上に、外的要因によるものなので、一つのURLで示されるウェブページに頼る方法では対処が難しい。そこで情報再構成部１５は再構成手順記憶部１８から予め記憶しておいた再構成手順を取得（ステップＳ６）し、その手順に従って抽出された複数の重複した情報要素を元に出力となる情報要素群を再構成する（ステップＳ７）。再構成手順には、各情報要素がどの目的とする情報に含まれているか、重複する情報要素の中で異なるデータが出現したときの対処の方法、そしてどのようなデータであれば信頼できるか等の情報を含んでいる。場合によっては情報再構成部１５が再構成手順に記録された方法によって再度ウェブページを取得して、新規に出力となる情報要素群を取得することも考えられる。

【００１３】このようにして構成された出力結果は出力部１４を介して出力され（ステップS８）、一連の処理を終了する。

【００１４】上記のように構成することによって、情報収集という目的における一部のウェブページの取得失敗や一部のウェブページに掲載された誤った情報に対する耐性を備えることが可能となる。このことによって出力する情報に欠損や誤りが生じる可能性を大きく低減させるとともに、情報の信頼性を高めることができる。

【００１５】既述した一連の動作は、定期的に繰り返されることによってリアルタイムに変化する情報、たとえば証券金融情報や気象情報などを最新情報に近い状態を維持しつつ常時、情報を提供するサービスにも資することもできる。

【００１６】次からは実際に取得するウェブページの例を示しながら説明する。ここに示す実施例は、ウェブページの作成言語として一般的に利用されているHTMLを例にとっている。本発明では抽出方法記憶部１７に記憶する抽出方法を適切に設計することで、ウェブページがXHTMLなどその他の形式で提供された目的とする情報であっても対処可能である。

【００１７】図３および図５は、１日のある時点の株価情報をウェブブラウザーで提供するために作成されたHTML文書の簡易な例を示している。そして図４および図６は、図３および図５を一般のウェブブラウザーによって表示させたときの表示イメージを表している。例では株価情報を提供するウェブサイトが２箇所あり、そのそれぞれが提供するHTML文書が図３および図５であるとしている。当然のこととして、図３および図５に示したHTML文書をウェブブラウザーで表示されると、図４および図６のように異なったものとなる。図３のものは、<TABLE..>～</TABLE>で囲まれた最初の領域に株価データが記録されているが、図５では２番目の<TABLE..>～</TABLE>領域に株価データが記録されている。

【００１８】このため図３のものと図５のものから株価情報を抽出するには、それぞれ異なる抽出方法を用いて行なわねばならない。これはそれぞれ個別の抽出方法が抽出方法記憶部１７に記憶されている必要があることを意味している。以降、説明のために図３のHTML文書を提供するウェブサーバをウェブサーバＡ、図５の方をウェブサーバＢと称する。

【００１９】また、ウェブサーバＡとウェブサーバＢ（表示画面はそれぞれ図４と図６）では、共通して保持されている情報と片方にしか存在しない情報とがある。銘柄名、現在値、取引時刻の３点は両方のウェブサーバから得ることができる。またウェブサーバＡからは前日終値が、ウェブサーバＢからは前日比が提供されるが、これは同一銘柄について見れば相互に補完が可能（前日比＝現在値－前日終値）であるから、相互に取得可能な情報要素と見ることができる。

【００２０】一方、ウェブサーバＡが提供する始値、高値、安値、およびウェブサーバＢが提供する出来高はそれぞれ片方にしか存在せず、また例示したような補完関係にもないため、片方のウェブサーバからのみ取得が可能な情報要素である。

【００２１】図７および図８は、それぞれウェブサーバＡおよびＢのそれぞれから株価情報を抽出するための抽出方法の一例を図示したものである。これらは抽出方法記憶部１７に記憶され、情報抽出部１３が情報要素を抽出する際に使用される。この例では情報抽出部１３は、ウェブサーバＡあるいはＢから与えられるHTML文書を、図７および図８に示した抽出手順に従って抽出していく。まず図７および図８に示すデータ開始位置から下方

に参照する。銘柄情報開始位置以降の銘柄の各データを、図７および図８の３～７行目に記載された位置のデータを書式に従って抜き出す。必要に応じて計算も行ないながら、データ終了位置まで、銘柄ごとに繰り返す。すると情報抽出部１３は銘柄名（Ａ社、Ｂ社、Ｃ社）とそれぞれの現在値、取引時刻、前日終値、前日比を取得できる。株価自体は共通のものであるから、原則としてウェブサーバＡとウェブサーバＢから得られるこれら共通に提供される情報要素は同一の値となるはずである。

【００２２】ここで各ウェブサーバが固有に提供するデータを取得したい場合は、高値、安値、出来高を抽出するための抽出方法を、図７および図８に示す抽出方法に書き加え抽出方法記憶部１７に記憶させればよい。さらにウェブサーバＡやウェブサーバＢ以外のウェブサーバＣにおいて高値、安値、出来高が提供されるのであれば、このウェブサーバＣの該当ウェブページを指し示すURLとそのウェブページについての抽出方法を、位置情報記憶部１１と抽出情報記憶部１７に加えることにより重複して取得することができるようになる。

【００２３】このようにして得られた重複した情報要素は、全て情報再構成部１５に送られる。そして情報再構成部１５はこれら重複した情報要素から出力となる１組の情報要素群を構成する。以下、取得時の情報要素のさまざまな取得状況において当該情報再構成部１５が取り得る、信頼性の高い情報要素群を再構成する再構成手順の事例を示す。

【００２４】情報再構成部１５が取得する複数の情報要素群は、本来同じ情報となるはずである。しかしながらさまざまな事情や不具合によって期待通りの結果とならない可能性がある。図９は株価情報を提供する９ヶ所のウェブサイトの提供するウェブページ１～９のそれぞれから得られた、Ａ社の株価の現在値を例示している。このうちウェブページ４ではウェブサイトの不具合かあるいはサービス停止など何らかの理由により、文書取得部１２によるウェブページの取得に失敗し情報が取得できなかったことを示している。またウェブページ２では、ウェブページそのものの取得には成功したものの、そこに含まれるＡ社の株価の現在値取得に失敗している。この原因としては、たとえばウェブサーバの動作不具合かウェブサイト運営者によってウェブページのデザインが変更されたことなどが考えられる。具体的には図６に示す画面表示例の、広告表示部分を削除した場合である。図６の元となるHTML文書である図５から、広告表示に相当する部分（図５（２））を削除したとすると、<TABLE..>～</TABLE>で囲まれた部分は図５（３）のみとなる。すると図８に示した情報抽出方法にある「２番目の<TABLE..>」は存在しないことになる。したがって、このような場合には情報要素の抽出に失敗するか、間違った情報を抽出してしまう。

【００２５】またウェブページ６からは「11:00」という株価を示す数値ではないデータが抽出されている。この原因としては、やはりウェブページのデザインが変更された可能性が考えられる。たとえば図５のウェブページを提供するウェブサイトが、当該ウェブページの上部に図６に示すウェブページと同様の表示を成す広告部分を追加したとする。すると図７に示す情報抽出方法における「１番目の<TABLE..>」は広告表示部分を指すことになり、広告を表す部分の範囲内の文字を株価情報とみなして抽出してしまう。ウェブページ４の場合、これと同様な理由で時刻表示部分をＡ社の株価とみなして抽出したものと想像できる。

【００２６】以上、ウェブページ２、４、６の各ケースでは株価情報が正常に抽出できなかったと判断され、最終的な出力には採用できない。したがってこれらは情報再構成部１５で出力結果の対象からは外される。

【００２７】次にウェブページ７からは「5,240,00」が、その他にも「¥1000」や「1000円」などの表現の異なる結果で抽出されたものがある。しかしながら後者は一般に表記され得る形の違いであり、Ａ社の株価が１０００円であることを表すウェブページが多数を占めている。よってこの図９の例からはＡ社の現在値の株価は１０００円とするのが妥当であると判断することができる。

【００２８】信頼性の高い情報要素の選定方法としては、上記のような多数決アルゴリズムが適用できる。情報源としては極力信頼のおけるウェブサイトを選定し、それぞれから目的とする情報を含むウェブページを複数取得するために位置情報を複数用意しておく。信頼性の高いウェブページばかりを取得できたとしても、既述の不具合を完全に排除することはできないが、より確からしい情報を得ることはできる。何らかの要因により正しい情報が取得できる可能性が低かったとしても、同時に複数のウェブページが同様の誤りを含むことは極めて少ないはずだからである。よって多数決の原理に従うことにより確率論的に十分確からしい情報要素を推定することができる。

【００２９】もっとも、信頼性の高い情報要素の方を選択できれば足りるため、多数決アルゴリズムに限定するものでもなくそれ以外の方法で行なってもかまわない。情報要素の信頼性を推定するさまざまな推論を使い分けるために、再構成手順記憶部１８に推定アルゴリズムを記述するようにしてもよい。このように構成すれば、収集する情報の種類や情報源となるウェブサイトの状況に応じた適切なアルゴリズムを適用することができる。

【００３０】次からはより信頼性のある情報要素の推定アルゴリズムの例について説明する。既述した多数決アルゴリズムは、信頼のおけるウェブページがある程度複数取得できなければならない。しかしながら現実には多数決アルゴリズムに必要な複数のウェブページが取得できない場合も有り得る。このような場合に有効と思われ

る推定アルゴリズムとして、ａ）制約条件による方法、ｂ）先着の情報を優先する方法、ｃ）値の域値による方法などが考えられる。

【００３１】ａ）制約条件による方法

目的とする株価情報が２箇所のウェブページより取得できると仮定する。また、これ以外のウェブページからは取得ができないものとする。一方のウェブページではＡ社株価の現在値が「１，０００円」、もう一方が「５，２４０円」という情報を提供していたとする。このような場合、どちらか一方の数値が実際に生じ得ないことが分かれば信頼性のある情報として他方の数値を採用することができる。生じ得ない情報を排斥することで、最終的な出力結果をより信頼性のあるものにすることが可能となる。

【００３２】たとえば、日本の株式市場においては多くの株式に対して「値幅制限」という制限が行なわれる。図１０は、この値幅制限の一例を示したものである。前日の終値が図１０の左側の条件を満たす株式では、前日終値から図１０の右側に示されている金額を超えて取引が行なわれないというものである。

【００３３】本発明の実施例では、再構成手順記憶部１８に図１０の制限値幅と、これが行なわれている株式を示す情報、および同条件を逸脱する株価情報は取得した株価情報を無効とする旨の条件を記憶しておくことによって行なえる。

【００３４】先の例でいえば、いまＡ社株式が値幅制限を受けている株式であり、Ａ社株式の前日終値が９２２円であったとする。再構成手順記憶部１８に記憶された上記条件を読み込んだ除法再構成部１５は、この株価が値幅制限に抵触するかどうかを判断する。情報再構成部１５は、抽出した情報要素がＡ社株式のものであることを判断し、前日終値と制限値幅表からＡ社株式の制限値幅が１００円であることを知得する。よってＡ社株式の取り得る現在値は８２２円～１，０２２円の間に無ければならないことが分かる。情報再構成部１５は抽出した「１，０００円」および「５，２４０円」の値を、上記値幅に照らし合わせ「５，２４０円」はおよそ有り得ない株価であると判断する。この判断を経て、情報再構成部１５は「１，０００円」という株価を確からしい情報要素として採用する。

【００３５】上記のように目的とする情報そのものに何らかの制約があることが事前に分かっているようなものについては、再構成手順記憶部１８に予め記憶しておくことによって誤った情報が混入することを避け、出力結果における情報の信頼性を高めることができる。

【００３６】目的とする情報によっては制約条件を重畳することにより、さらに信頼性を高めることが可能である。たとえば株価情報の場合、１日の通算出来高は減少しない旨を制約条件として記憶しておくことにより、出来高の情報抽出における不具合を排除できる。これ以外にも現実的には生じ得ない、あるいは極めて低い確率でしか生じ得ないと考えられる情報を検出するように条件を組み立てることも、信頼性の高い情報を構成するには有効である。先の株価情報の例でいえば、たとえば値幅制限を持たない一部の株式で、図１０に示したような値幅制限の１０倍を越える値幅の変動が見つかった場合、当該株価情報を採用しないというようにである。

【００３７】これら情報要素の選別を行なってもなお、必要とする信頼性のある情報を取得することが困難な場合も考えられる。必要に応じて、前述した多数決アルゴリズムと併用するなど別種の判定アルゴリズムをさらに組み合わせることで効果を高めても良い。

【００３８】あるいは出力結果に極めて高い信頼性を要求される場合にあっては、判定に十分な情報が収集できなかった情報に関して「取得不可」であった旨の出力をしても良い。たとえば多数決アルゴリズム併用時において取得した情報の中で全てが異なる情報を示し多数を占める値が認められない場合や、得た情報が全て制約条件範囲を逸脱するような場合である。

【００３９】ｂ）先着の情報を優先する方法

情報提供の要請から、よりリアルタイム性を確保するために取得した情報の取得順により採用する情報要素を決定する方法である。ウェブサイトから情報を取得する場合を考えたときに、ウェブサイト毎に応答速度が異なる場合が多い。あるいは不測の事態により、特定のウェブサイトがサービスを停止している場合も有り得る。このような場合に複数の目的とする情報が収集されるのを待っていては必要とするリアルタイム性を満たせないことがある。複数ある位置情報のウェブサイトから、最先に得られた目的とする情報を優先して採用する（先着優先アルゴリズム）ものである。再構成手順記憶部１８が記憶する推定アルゴリズムには、最先に得られた目的とする情報を採用する旨の条件を記憶しておくことで選定がされる。

【００４０】ｃ）値の域値による方法

取得した複数の情報のいずれかを選択する以外に、得られた複数の情報を統計的に総合して得られた結果を採用する方法も考えられる。たとえば目的とする情報が数値であった場合、複数の情報の平均値を出力とする（平均値アルゴリズム）、あるいは最大値／最小値を出力とする（最大値／最小値アルゴリズム）、などである。さまざまな統計的手法が考えられるが、情報の性質や得られる情報の活用場面を考慮して決定される。前出と同様、再構成手順記憶部１８には平均／最大／最小、その他の統計的手法を用いて出力結果を構成する旨が記憶される。

【００４１】今までの説明では個々に得られた複数の情報を一つの集合とみなして、その中から信頼性の高いものを推定するための方法を示してきた。次からはこれら情報を提供するウェブサイトあるいはウェブページの信

頼性を考慮して、より信頼性の高い情報を構成する方法を説明する。

【００４２】あるウェブページの取得に失敗した時、それがウェブサイトの一時的な不具合であればその復旧を待つことで、再び情報源として使用できる。しかしながらそれが一時的な要因によらない場合、たとえばウェブサービスそのものの停止やURL等の位置情報、ウェブページの構成などが予告無く変更された場合は、変更された内容によって位置情報や抽出方法の更新作業が必要となる。本来、情報源として使用することが難しいと判断される事態が判明したときは、本発明にかかる情報収集装置の管理者は当該ウェブページの取得を停止し、新たなウェブページを選定し取得先として加えるなどの管理を行なうべきである。しかしながら上記管理者が上記措置を講ずる前にも、出力として得られる情報の信頼性を維持できる仕組みが備えられていることが望ましい。

【００４３】また目的とする情報の物理的な取得に失敗する以外にも、抽出した情報の一部が既述したような制約条件を満たしていないことを検出した場合には、当該位置情報から得られる目的とする情報の信頼性は全体として著しく低下したと判断すべきである。なぜなら一部に許容できない誤りがある場合、その全体の信頼性も低下していると考えられるからである。たとえばウェブサイトが提供するウェブページに広告目的で<TABLE..>～</TABLE>のような情報を付加する例を示したが、この例のように現行使用している抽出条件が適用できない危険性が極めて高い。たとえばHTMLで書かれた文書では、一部の変更の影響がHTML文書全体に及び易い。そのため抽出された情報が正しいように見えても、それは目的とする情報要素外の別の情報かも知れないからである。

【００４４】このような事態を想定すると、位置情報で示されたウェブページ毎に信頼性に基づく順列を付与し、この順列に従ってそれぞれから得られる情報を差別化する必要がある。上記の順列は位置情報と共に位置情報記憶部１１に記憶され、再構成手順記憶部１８に記憶された手順に従って情報再構成部１５によって読み出され利用される。同アルゴリズムは前出の各アルゴリズムと共に適用しても良く、そのように構成すると出力される情報の信頼性をより高めることができる。

【００４５】図１１は、ウェブページ１～５を提供するウェブサイトがあるときのそれぞれの抽出結果を示す図である。これは一例としての、各ウェブページの信頼性を示す順列情報に基づき情報再構成１５の内部に格納されている状況を示している。上方に現れているウェブページほど信頼性が高いことを意味している。

【００４６】図１１ではウェブページ１および３に情報取得時に不具合が発生している。一度の情報取得失敗は、以降の情報取得にとって信頼性を損なったと判断する。このため情報再構成部１５は図１２に示すように、各ウェブページの信頼性を示す順序を入れ替える。このときウェブページ３は情報抽出には成功しているものの制約条件を満たしていない情報が存在するため、最も信頼性の低い位置に入れ替えられている。先に説明したように、一部に問題を持つウェブページは全体的に問題を含んでいる可能性が高いと考えられるためである。場合によっては当該ウェブページの取得を中止するようにしても良い。

【００４７】ウェブページ１は取得に失敗したために信頼性の順位を下げられているが、比較的短い時間に回復が見られれば信頼性が回復され、順列が戻されるようにしても良い。このとき一向に回復が見られない（たとえば数分～数時間）ようであれば当該ウェブページのサービスの停止あるいはURLなどの位置情報が変更されたと判断し、以降の当該ウェブページの取得を中止するようにしても良い。取得を中止したウェブページは、本発明にかかる情報取得装置の管理者等が判断し必要ならば抽出条件や位置情報を更新した後に、再び信頼性を回復させる。

【００４８】ウェブページ毎に信頼性を与えることができるようになると、さらに柔軟な情報取得が行なえる。たとえば各ウェブページの信頼性が図１２で示される状況にあるとき、情報再構成部１５は要求発行部１６に指示し信頼性の高いウェブページ２、４、５のみを取得させる。このときウェブページ２、４、５から信頼性に足らない情報しか取得できなかったときに限り、ウェブページ１を取得するように構成することができる。あるいは繰り返し情報を取得するような場合には、信頼性の高いウェブページ２，４，５の取得頻度を高め、逆にウェブページ１のそれを低くするなどの差別化を図ることも可能である。

【００４９】次に上述した信頼性を示す情報と、既述の多数決アルゴリズムとを併用したときの信頼性の高い情報の判定方法の例を示す。具体的には信頼性の順序が高い順にウェブページを取得し、その取得情報の中で同一のものが過半数を超えた時点で決する方法、あるいはウェブページの信頼性を数値換算し、そこから得られる情報にこの換算値を加点して行く方法とが考えられる。前者を（１）優先多数決アルゴリズム、後者を（２）議決権アルゴリズム、また両者を組み合わせたものを（３）優先議決権アルゴリズムと呼ぶ。

【００５０】（１）優先多数決アルゴリズム

信頼性の判断が図１２のような順序にある場合に、最上位のウェブページ（ウェブページ２）から順に取得する。情報抽出部１３はそのウェブページに含まれる情報要素を抽出し、抽出した情報要素の値を蓄積する。次に２番目にあるウェブページ（ウェブページ４）から同様に情報要素を抽出し、これを蓄積する。このとき、全取得対象であるウェブページ数に占める、同じ情報要素の値が過半数を満たす数に到達するまで順次取得してゆく。例ではまず、ウェブページの２、４、５を要求し、

これらから抽出した情報要素の値が一致するならば多数決理論からその値を採用する。この時点で信頼性のある情報が確定できれば、残りのウェブページ１および３を取得する必要がない。つまりこのように判断することにより、信頼性のある情報の確定のために費やす時間を短縮できる可能性がある。仮にこの時点で過半数が得られなかった場合でも、順次つぎの信頼性を与えられているウェブページを取得し、多数決が取られた時点で終了するようにすれば良い。いずれとしても極力全てのウェブページを取得せずとも確からしい値の確定が行なえる。また信頼性のあるウェブページから取得して行くため、判定の初期から信頼性が高いと思われる情報を使用できる。このことからも出力結果における信頼性を確保できる。

【００５１】場合によっては多数決に必要な数のウェブページを取得できないことも考えられる。そういった場合には、適宜過半数に必要な採択数を変更しても良い。たとえばウェブページ１、４の取得に失敗したとき、ウェブページ２、４、５の３箇所から得られる情報要素のみを用いて決すれば良い。このとき過半数は２票として判定される。

【００５２】取得可能なウェブページを全て取得しても、いずれの情報要素の値に対しても過半数が得られない場合も考えられる。そのような場合には、多数を得た情報要素の値とするか、取得できた内の最も信頼性の高いウェブページから得られた情報要素の値とするか、あるいはそれらの総合判断で決定すれば良い。

【００５３】（２）議決権アルゴリズム

図１２に示すような信頼性順序を数値換算し、これをいわゆる議決数として多数決を取る。優先多数決アルゴリズムでは取得順に信頼性を反映させるが、本方法では投票時に重み付けを行なう点で異なる。たとえば図１２の例においては信頼性順序の数値換算したものをそれぞれ「５、４、３、２、１」とすることができる。ウェブページ１～５のそれぞれから得られる情報要素の示す値に、図１２に示す順列に従ってそれぞれ「２、５、１、４、３」の議決権を割り当てる。そして情報要素の示す値の一致するものに議決権を加算し、最も高い得票数を得た情報要素の値を確からしい値として採用する。このように議決権の考え方を導入することによって、より信頼性のある値を選定することが可能となる。

【００５４】（３）優先議決権アルゴリズム

優先多数決アルゴリズムと議決権アルゴリズムを組み合わせて適用する方法である。優先多数決アルゴリズムのときと同様に、情報再構成部１５は要求発行部１６に指示して、信頼性の高いウェブページから順に取得していく。同時に議決権アルゴリズムのときのように、位置情報の指し示すそれぞれのウェブページに議決権を割り当てておく。そして得られた情報要素の値に一致するものに対し、当該ウェブページに割り当てられた議決権を加算してゆく。このとき上記のように議決権を加算した後の得票数が、全得票数の過半数を獲得するに至ったかどうかを加算の都度順次判定する。特定の情報要素の値に対する得票数が過半数を越えた時点で、当該値を確からしい値として採用し、順次行なっていたウェブページの取得を終了する。

【００５５】図１２の例では、それぞれの信頼性順列に「５、４、３、２、１」を付与すると、全得票数は１５票となる。ウェブページ２の情報要素の値が「Ａ」であったとき、該「Ａ」に対して５票が加算される。次のウェブページ４もまた「Ａ」であった場合、さらに４票が加算される。このとき「Ａ」に対する得票数は９票となり、全得票数「１５票」における過半数を獲得したことになる。最短の場合、この時点で確からしい情報要素の値の判定が終了する。

【００５６】上述したように、信頼性を確保したまま優先多数決アルゴリズムに比べてもより短時間で判定を終了することが可能である。

【００５７】また位置情報が指し示すウェブページの信頼性を数値化し、取得の度に順列の変更が成されることにより、本発明にかかる情報収集装置の管理者にも有益である。取得の際に不具合が多く発生するウェブページは、信頼性を示す順列の下位に配置される。よって管理者は、当該ウェブページを指し示す位置情報を変更する、または情報要素の抽出方法を変更するなど、適切な措置を講ずることかできる。このことにより当該情報収集装置から得られる出力結果の信頼性を維持することが可能である。

【００５８】上記のようなさまざまな方法により、より信頼性のある情報要素を選定することで、ある時点におけるより高い信頼性を持つ情報を選定し、それを再構成して出力することができる。出力の方法は、たとえば出力結果が株価情報であればグラフ化するなどして画面に表示することも可能である。あるいはファイル等に蓄積し、データとして二次利用に供することもできる。

【００５９】また本発明にかかる情報収集装置では、信頼性の高い情報の選定を短時間に行なうことが可能である。これは多数決アルゴリズムのような複数の情報をつき合せるアルゴリズムを採用すれば、統計的な計算を必要とすることもなく、あるいは一定時間経過後の前後に渡る値を採取する必要もないからである。よってリアルタイムに極めて近い情報を提供する情報収集装置を得ることも可能となる。

【００６０】

【発明の効果】本発明によれば、インターネット等に一般に公開されている情報を元に、信頼性の高い情報を収集し、加えて二次利用も可能な形式の情報を得られる情報収集が可能となる。

【図面の簡単な説明】

【図１】本発明の一実施形態における情報収集装置の構

成図の一例を示す図である

【図２】本発明の一実施形態における情報収集装置の動作フローの一例を示す図である。

【図３】株価情報を提供するウェブページのHTML文書の一例を示す図である。

【図４】図３に示すHTML文書の表示イメージの一例を示す図である。

【図５】株価情報を提供するウェブページのHTML文書の一例を示す図である。

【図６】図５に示すHTML文書の表示イメージの一例を示す図である。

【図７】本発明の一実施形態における、図３に示すHTML文書から情報要素を抽出する手順の一例を示す図である。

【図８】本発明の一実施形態における、図５に示すHTML文書から情報要素を抽出する手順の一例を示す図である。

【図９】複数のウェブページから得られた情報要素の取得状況の一例を示す図である。

【図１０】株式市場で採用されている、株価制限値幅表の一例を示す図である。

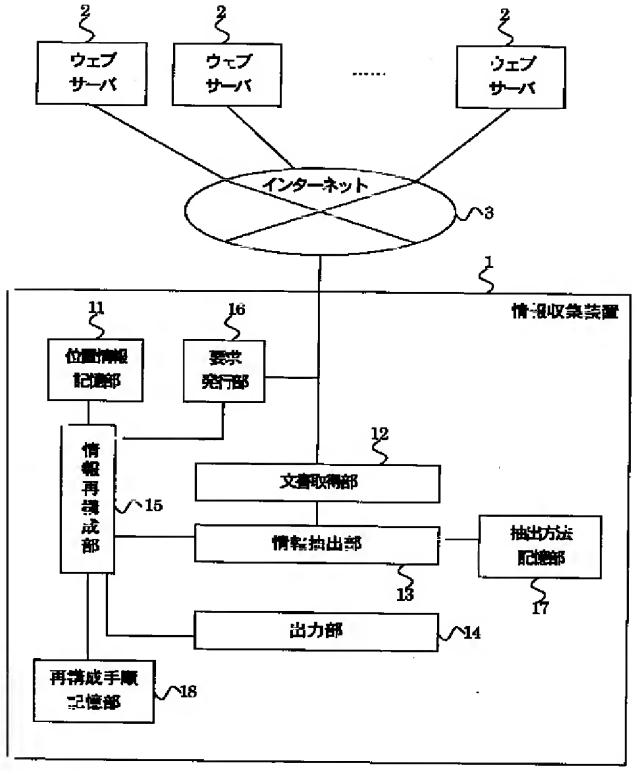
【図１１】本発明の一実施形態における取得ウェブページの信頼性情報と情報抽出状況の一例を示す図である。

【図１２】本発明の一実施形態における、情報抽出状況に従って並び替えた取得ウェブページの信頼性情報と情報抽出状況を示す図である。

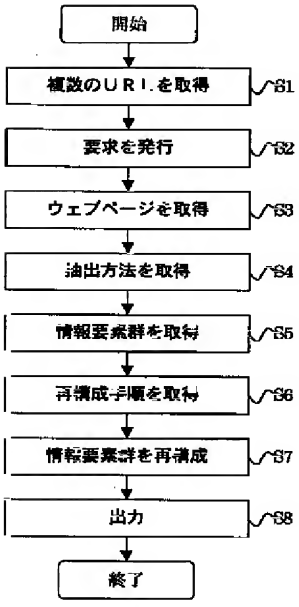
【符号の説明】

１　情報収集装置
２　ウェブサーバ
３　インターネット
１１　位置情報記憶部
１２　文書取得部
１３　情報抽出部
１４　出力部
１５　情報再構成部
１６　要求発行部
１７　抽出方法記憶部
１８　再構成手順記憶部

【図１】



【図２】



【図１０】

前日終値	制限値幅
100円未満	30円
200円未満	50円
500円未満	80円
1,000円未満	100円
1,500円未満	200円
…	…

【図４】

銘柄名	始値	高値	安値	現在値	取引時刻	前日終値
A社	942	1034	930	1000	11:00	922
B社	54000	55000	49000	50000	12:35	56000
C社	49	51	49	50	10:42	49

【図3】

```
<HTML>
<HEAD><TITLE>株価情報</TITLE></HEAD>
<BODY>
<TABLE BORDER=1 CELLSPACING=0>
<TR><TH>銘柄名</TH><TH>始値</TH><TH>高値</TH><TH>安値</TH><TH>現在値</TH>
<TH>取引時刻</TH><TH>前日終値</TH></TR>
<TR><TD>A社</TD><TD>942</TD><TD>1034</TD><TD>930</TD><TD>1000</TD>
<TD>11:00</TD><TD>922</TD></TR>
<TR><TD>B社</TD><TD>54000</TD><TD>55000</TD><TD>49000</TD><TD>50000</TD>
<TD>12:35</TD><TD>56000</TD></TR>
<TR><TD>C社</TD><TD>49</TD><TD>51</TD><TD>49</TD><TD>50</TD>
<TD>10:42</TD><TD>48</TD></TR>
</TABLE>
</BODY>
</HTML>
```

①

【図5】

```
<HTML>
<HEAD><TITLE>現在の株価</TITLE></HEAD>
<BODY>
<TABLE>
<TR><TD><IMG SRC="images/ad1.jpg"></TD><TD><IMG SRC="images/ad2.jpg"></TD>
<TD><IMG SRC="images/ad3.jpg"></TD></TR>
</TABLE>
<TABLE BORDER=1 CELLSPACING=0>
<TR><TH>銘柄名</TH><TH>現在値</TH><TH>取引時刻</TH>
<TH>前日比</TH><TH>出来高</TH></TR>
<TR ALIGN=RIGHT><TD ALIGN=LEFT>A社
</TD><TD>1,000</TD><TD>11:00</TD><TD>+78</TD><TD>5,240,000</TD></TR>
<TR ALIGN=RIGHT><TD ALIGN=LEFT>B社
</TD><TD>50,000</TD><TD>12:35</TD><TD>-6,000</TD><TD>123</TD></TR>
<TR ALIGN=RIGHT><TD ALIGN=LEFT>C社
</TD><TD>50</TD><TD>10:42</TD><TD>+1</TD><TD>3,000</TD></TR>
</TABLE>
</BODY>
```

②

③

【図6】

広告1　広告2　広告3

銘柄名	現在値	取引時刻	前日比	出来高
A社	1,000	11:00	+78	5,240,000
B社	50,000	12:35	-6,000	123
C社	50	10:42	+1	3,000

【図7】

データ開始位置	1番目の<TABLE..>の次の</TR>の次
銘柄情報開始位置	次の<TR>の次
銘柄名	次の<TD>と</TD>の間
現在値	4番目の<TD>と</TD>の間
取引時刻	次の<TD>と</TD>の間
前日終値	次の<TD>と</TD>の間
前日比	現在値ー前日終値
データ終了位置	次の</TABLE>の位置

【図8】

データ開始位置	2番目の<TABLE..>の次の</TR>の次
銘柄情報開始位置	次の<TR..>の次
銘柄名	次の<TD..>と</TD>の間
現在値	次の<TD>と</TD>の間のカンマ区切り数値
取引時刻	次の<TD>と</TD>の間
前日比	次の<TD>と</TD>の間のカンマ区切り符号つき数値
前日終値	現在値ー前日比
データ終了位置	次の</TABLE>の位置

【図9】

	A社の株価の現在値
ウェブページ1	1000
ウェブページ2	情報抽出失敗
ウェブページ3	1,000
ウェブページ4	ウェブページ取得失敗
ウェブページ5	1000円
ウェブページ6	11:00
ウェブページ7	5,240,000
ウェブページ8	¥1000
ウェブページ9	1,000円

【図11】

	取得および抽出状況
ウェブページ1	ウェブページ取得失敗
ウェブページ2	正常に抽出
ウェブページ3	抽出情報の一部が制約条件を満たさない
ウェブページ4	正常に抽出
ウェブページ5	正常に抽出

【図12】

	取得および抽出状況
ウェブページ2	正常に抽出
ウェブページ4	正常に抽出
ウェブページ5	正常に抽出
ウェブページ1	ウェブページ取得失敗
ウェブページ3	抽出情報の一部が制約条件を満たさない